Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Estatística

Dissertação de Mestrado

A DISTRIBUIÇÃO TOUCHARD
E SUAS APLICAÇÕES

por
Sandro Barbosa de Oliveira

Orientador: Prof. Dr. Raul Yukihiro Matsushita

Brasília - Distrito Federal
Dezembro de 2016

SANDRO BARBOSA DE OLIVEIRA
sandrobarboliveira@gmail.com

A DISTRIBUIÇÃO TOUCHARD
E SUAS APLICAÇÕES

Dissertação apresentada ao
Departamento de Estatística do
Instituto de Ciências Exatas da
Universidade de Brasília como
requisito parcial à obtenção do
título de Mestre em Estatística.

Universidade de Brasília
Brasília, Dezembro de 2016

TERMO DE APROVAÇÃO

Sandro Barbosa de Oliveira

A DISTRIBUIÇÃO TOUCHARD
E SUAS APLICAÇÕES

Dissertação apresentada ao Departamento de Estatística do Instituto de Ciências Exatas da Universidade de Brasília como requisito parcial à obtenção do título de Mestre em Estatística.

Data da defesa: 19 de dezembro de 2016

Orientador:

Prof. Dr. Raul Yukihiro Matsushita
Departamento de Estatística, UnB

Comissão Examinadora:

Prof. Dr. Bernardo Borba de Andrade
Departamento de Estatística, UnB

Prof. Dr. Pushpa Narayan Rathie
Departamento de Estatística e Matemática Aplicada, UFC

Agradecimentos

Agradeço à minha família: Vivi, Bela e Maria, meus pais e meus irmãos.

Ao professor Raul pela orientação e confiança.

Aos professores Bernardo, Nakano, Cira, Alan, Antônio Eduardo e Gilardoni pelo conhecimento transmitido, fundamental para a construção desse trabalho.

Aos amigos que, direta ou indiretamente, contribuiram para a conclusão desse trabalho.

Finalmente, agradeço à Rede SARAH pelo apoio e à Uiversidade de Brasília pela oportunidade de concluir mais uma etapa da minha formação acadêmica.

"*It's the job that's never started as takes longest to finish.*"

J.R.R. Tolkien

Conteúdo

Lista de Figuras

Lista de Tabelas

Resumo

A distribuição de Poisson é uma das mais importantes distribuições de probabilidade, sendo amplamente utilizada para modelagem de dados provenientes de experimentos de contagem. Seu único parâmetro é também sua média e sua variância, o que a torna inadequada para a modelagem de dados com subdispersão, superdispersão e excesso de zeros.

Nesta dissertação será apresentada a distribuição Touchard, uma generalização com dois parâmetros da Poisson, com a proposta de modelar dados com subdispersão, superdispersão e excesso de zeros. Será também introduzido o modelo de regressão Touchard e uma generalização com três parâmetros.

Diversas aplicações ilustraram como a distribuição Touchard pode ser uma alternativa competitiva para modelagem de dados não-Poisson, equiparando-se com as mais clássicas e recentes generalizações da Poisson.

Palavras-chave *Distribuição de Poisson Generalizada; Superdispersão; Distribuição de Poisson; Distribuição Touchard; Subdispersão; Distribuição Inflada com Zeros; Regressão Touchard.*

Abstract

The Poisson distribution, one of the most important distributions in probability theory, has been widely used to model count data. The Poisson distribution depends on a single parameter λ. The expected value and variance of a Poisson-distributed random variable are both equal to λ, so using standard Poisson model with under or overdispersed data may result in lack-of-fit.

This dissertation presents a two-parameter extension of the Poisson distribution: the Touchard distribution. It is a flexible distribution that can account for both under- or overdispersion and concentration of zeros that are frequently found in non-Poisson count data. Touchard regression and three-parameter extension of the Poisson distribution will also be shown in this work.

Several applications will illustrate the capabilities of this approach to be a useful model for assessing non-Poisson data.

Keywords *Generalized Poisson distribution; Overdispersion; Poisson distribution; Touchard distribution; Underdispersion; Zero-inflated distribution; Touchard regression.*

Introdução

Em diversas áreas do conhecimento, muitos experimentos ocorrem com a observação da contagem de eventos em determinado tempo ou espaço. Em geral, se os eventos aleatórios ocorrem de forma independente a uma taxa (tempo) ou densidade (espaço) constantes, então a contagem desses eventos por unidade de tempo ou área segue a distribuição de Poisson.

Na distribuição de Poisson seu único parâmetro λ é tanto sua média como sua variância. Há situações nas quais os dados indicam subdispersão (a média é maior que a variância), superdispersão (a média é menor que a variância) ou, ainda, excesso de zeros (por exemplo, contagem de eventos raros).

Dados com características de subdispersão, superdispersão e excesso de zeros podem gerar vieses nos erros padrões das estimativas de λ se a distribuição de Poisson for utilizada.

É possível encontrar na literatura algumas generalizações da Poisson para modelagem de dados de contagem em que a média difere da variância. Destacam-se as distribuições Binomial Negativa, Poisson Generalizada [1], Poisson Dupla [2, 3] e Conway-Maxwell-Poisson [4, 5].

Algumas publicações mais recentes também têm proposto alternativas para a modelagem de dados do tipo não-Poisson, como as distribuições Nova Generalização Poisson-Lindley [6] e a Conway-Maxwell-Poisson Estendida [7].

Há, ainda, modelos para dados de contagem com excesso de zeros, chamados Poisson inflado com zeros [8].

Como uma alternativa viável e simples para modelagem de dados não-Poisson, nessa dissertação será apresentada a distribuição Touchard [9], uma generalização da distribuição de Poisson com dois parâmetros que permite modelar, não apenas dados com subdispersão ou superdispersão, mas também dados com excesso de zeros. Essa dissertação está organizada conforme descrito nos parágrafos a seguir.

No Capítulo 1, Seção 1.1, é feita uma breve introdução à distribuição Poisson, trazendo um pouco da história, propriedades e publicações clássicas dessa importante distribuição. Essa introdução certamente é dispensável para a maioria dos leitores se considerarmos a importância da distribuição de Poisson na teoria de Probabilidade e Estatística. De qualquer forma, a opção de manter esse capítulo nesta dissertação visa valorizar a história da Estatística. Na Seção 1.2 revisamos a regressão de Poisson no contexto dos modelos lineares generalizados. A Seção 1.3 traz uma breve revisão sobre variações da Poisson que se propõem a contornar o problema de superdispersão e/ou subdispersão em dados de contagem.

A nova generalização da Poisson será introduzida no Capítulo 2, onde serão apresentadas as propriedades da distribuição Touchard e métodos numéricos para estimação de seus parâmetros pelo método de máxima verossimilhança. O Capítulo 3 traz a regressão Touchard, um modelo linear generalizado no qual a variável resposta Y tem distribuição Touchard(λ, δ) e seus parâmetros λ e δ podem ser explicados por variáveis independentes. Os detalhes da implementação computacional poderão ser vistos no Capítulo 4.

Uma extensão da distribuição Touchard com três parâmetros será exposta no Capítulo 5. A adição de um terceiro parâmetro tem como objetivo melhorar o ajuste a dados com concentração de zeros.

O desempenho da distribuição Touchard em relação à distribuição de Poisson e a outras alternativas para dados não-Poisson será mostrado mais a frente no Capítulo 6, a partir de ilustrações com dados de problemas reais e de relevância científica.

Capítulo 1

Modelos para dados de contagem

1.1. Distribuição de Poisson

A distribuição de Poisson é uma das mais importantes distribuições discretas da teoria de Probabilidade e Estatística. Nas palavras de Sir Ronald Fisher, "*among discontinuous distributions, the Poisson series is of first importance*". [10]

A distribuição foi introduzida por Siméon Denis Poisson (1781-1840) e publicada em 1837, nos últimos anos de sua vida, no trabalho "Recherches sur la probabilité des jugements en matière criminelle et en matière civile" (Pesquisa em Probabilidade sobre Julgamentos nas Matérias Penal e Civil, em tradução livre), como resultado do seu interesse na aplicação da probabilidade na administração da justiça. [11]

Ladislaus von Bortkiewicz foi quem primeiro identificou uma aplicação prática da distribuição de Poisson, quando reconheceu uma relação entre a fórmula da Poisson e alguns tipos de dados discretos. Seu exemplo mais conhecido investigou o número de soldados do exército da Prússia mortos acidentalmente por coices de cavalos.

A partir de 1909 a distribuição Poisson começou a ser conhecida no meio científico, aparecendo com a nomenclatura atual "distribuição de Poisson". Desde então, diversas áreas do conhecimento têm utilizado a distribuição de Poisson em suas aplicações, como a indústria, agricultura e ecologia, biologia, medicina, telefonia, acidentes, comércio, teoria das filas, entre outros.

Uma variável aleatória discreta X é dita ter distribuição de Poisson com parâmetro $\lambda > 0$ se, para $k = 0, 1, 2, ...$, a função de probabilidade de X é dada por

$$P(X = k) = f(k; \lambda) = \frac{\lambda^k e^{-\lambda}}{k!},$$

e sua função geratriz de momentos pode ser escrita na forma

$$M_X(t) = \mathrm{E}[e^{tX}] = \sum_{k=0}^{\infty} \frac{e^{tk} e^{-\lambda} \lambda^k}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{(\lambda e^t)^k}{k!} = e^{-\lambda} e^{\lambda e^t} = e^{\lambda[e^t - 1]}.$$

O parâmetro λ é igual à esperança de X e também à sua variância:

$$\mathrm{E}[X] = \mathrm{Var}[X] = \lambda$$

Essa propriedade, que faz a distribuição de Poisson ser brilhantemente simples, pode ser vista como uma limitação para dados de contagem em que, claramente, tem-se a média da distribuição menor ou maior que a sua variância.

1.2. Regressão de Poisson

A regressão de Poisson faz parte de uma classe de modelos de regressão linear chamada de Modelos Lineares Generalizados (GLM, do inglês *Generalized Linear Models*). Propostos originalmente por Nelder e Wedderburn [12], os modelos lineares generalizados são uma extensão dos clássicos modelos de regressão linear.

Um GLM consiste em três componentes:

- Um componente aleatório Y_i chamado variável resposta. No texto original, a distribuição de Y pertence à família exponencial.
- Um preditor linear, que é uma função linear das variáveis preditoras (ou regressores),

$$\eta_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_2 X_{i3} + \ldots + \beta_k X_{ik}.$$

- Uma função suave e invertível $g(\cdot)$, chamada função de ligação, que transforma a esperança da variável resposta $\mu_i \equiv \mathrm{E}(Y_i)$ no preditor linear,

$$g(\mu_i) = \eta_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_2 X_{i3} + \ldots + \beta_k X_{ik}.$$

  Como a função de ligação é invertível, podemos escrever

$$\mu_i = g^{-1}(\eta_i) = g^{-1}(\alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_2 X_{i3} + \ldots + \beta_k X_{ik}).$$

Para dados de contagem, é assumido que Y tem distribuição de Poisson e a função de ligação é o logaritmo neperiano. Assim, a regressão de Poisson pode ser

escrita na forma

$$\log_e(\mathrm{E}[Y_i]) = \eta_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_2 X_{i3} + \ldots + \beta_k X_{ik},$$

sendo

$$\mathrm{E}[Y_i] = e^{\eta_i} = \exp\{\alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_2 X_{i3} + \ldots + \beta_k X_{ik}\}.$$

Para outras distribuições da família exponencial, outras funções *link* podem ser utilizadas, como ilustrado na Tabela 1.2.1.

Tabela 1.2.1: Função de ligação canônica, domínio da variável resposta e função de variância condicional para famílias exponenciais

Família	*Link*	Amplitude de Y_i	$\mathrm{Var}(Y_i\|\eta_i)$
Normal	Identidade	$(-\infty, +\infty)$	ϕ
Binomial	Logit	$\frac{0,1,\ldots,n_i}{n_i}$	$\frac{\mu_i(1-\mu_i)}{n_i}$
Poisson	Log	$0, 1, 2, \ldots$	μ_i
Gama	Inversa	$(0, +\infty)$	$\phi\mu_i^2$

Nota: ϕ é o parâmetro de dispersão, η_i é o preditor linear, μ_i é o valor esperado da variável resposta Y_i e n_i é o número de ensaios da distribuição Binomial.

1.3. Generalizações da Poisson

Há na literatura várias generalizações da distribuição de Poisson. Essas generalizações buscam adequar a clássica distribuição a dados de contagem do tipo não-Poisson, contornando os problemas de subdispersão, superdispersão e/ou excesso de zeros nos dados.

Define-se subdispersão quando os dados apresentam variância menor que a média. Por outro lado, nos casos em que a variância dos dados é maior que a média, tem-se superdispersão. O excesso de zeros na distribuição dos dados também torna o modelo de Poisson inadequado.

Já em 1919, William Gosset, sob o pseudônimo de "Student", tentava modificar a Poisson para permitir a aplicação em dados com média e variâncias desiguais [13], o que foi alcançado por Greenwood e Yule [14] com uma mistura de distribuições. Ainda no contexto de mistura de distribuições, ao considerar que o parâmetro λ tem distribuição Gama, o resultado obtido é hoje chamado Pólya-Eggenberger [15]. A distribuição Binomial Negativa tem sido utilizada como alternativa à distribuição Poisson para os casos de dados com superdispersão, permitindo uma maior flexibilidade na relação média/variância [16].

Ao longo dos anos surgiram outras propostas de generalização da Poisson, como, por exemplo, a *Generalized Poisson Distribution* proposta por Satterthwaite [17], a *compound Poisson distribution* publicada por Maceda [18] e a Poisson-Lindley [19].

Além disso, outras generalizações da Poisson trouxeram como proposta a adição de um segundo parâmetro que permita o melhor ajuste de dados com sub ou superdispersão. Nesse sentido, podemos citar as distribuições Conway-Maxwell-Poisson [4] - bastante explorada por Shmueli et al. [5] -, Poisson Generalizada [1]

e Poisson Dupla [2].

Para os casos de dados com excesso de zero, Lambert [8] propôs o modelo de regressão de Poisson inflado com zeros (ZIP - zero-inflated Poisson model).

Apesar das diversas propostas de generalização da Poisson, novas distribuições continuam surgindo com o proposito de modelar dados com sub e superdispersão. Mais recentemente, Bhati et al. [6] propuseram a Nova Poisson-Lindley Generalizada – *New Generalized Poisson-Lindley Distribution* (NGPL), uma versão com três parâmetros da distribuição generalizada Lindley. Ainda em 2016, Chakraborty e Imoto [7] publicaram a Conway-Maxwell-Poisson Estendida – *Extended Conway-Maxwell-Poisson distribution* (ECOMP), uma versão com quatro parâmetros da distribuição Conway-Maxwell-Poisson (COMP).

O desempenho da generalização da distribuição de Poisson proposta nesta dissertação será comparado com algumas das distribuições acima. A tabela a seguir traz um indicador da popularidade das principais generalizações da Poisson e de recentes publicações. A estatística de citações foi baseada no índice de citações do Google Acadêmico (*Google Scholar*) em pesquisa na data 6 de novembro de 2016.

Tabela 1.3.1: Número de citações das generalizações da Poisson.

Distribuição	**Autor(es)**	**Citações**
Binomial Negativa	Bliss & Fisher (1953)	1.147
Poisson Generalizada	Consul & Jain (1973)	385
Poisson Dupla	Efron (1986)	393
Conway-Maxwell-Poisson	Conway & Maxwell (1961)	155
Poisson Inflada com Zeros	Lambert (1992)	2.436
Nova Poisson-Lindley Generalizada	Bhati et al. (2015)	1
Conway-Maxwell-Poisson Estendida	Chakraborty & Imoto (2016)	0

Fonte: Google Acadêmico (pesquisa em 06/11/2016).

As seções a seguir trazem um breve resumo dessas generalizações da Poisson. As letras gregas utilizadas para representar os parâmetros das distribuições podem se repetir. Vale ressaltar, portanto, que a cada seção esses parâmetros serão referentes

exclusivamente à distribuição em questão.

1.3.1. Binomial Negativa

O modelo Binomial Negativo, também chamado de Poisson-Gama, é uma generalização do modelo Poisson que permite modelar dados com superdispersão. Esse modelo assume que o parâmetro λ tem distribuição Gama(α, β).

Seja $X|\Lambda$ uma variável aleatória com distribuição de Poisson e $\Lambda \sim$ Gama(α, β), então

$$\begin{aligned} P(X = x) &= \frac{1}{\Gamma(\alpha)\beta^\alpha} \int_0^\infty \frac{e^{-\lambda}\lambda^x}{x!} \lambda^{\alpha-1} e^{-\lambda/\beta} d\lambda \\ &= \frac{1}{x!\Gamma(\alpha)\beta^\alpha} \int_0^\infty \lambda^{\alpha+x-1} e^{-\lambda(1+1/\beta)} d\lambda \\ &= \frac{1}{\Gamma(x+1)\Gamma(\alpha)\beta^\alpha} \Gamma(\alpha+x) \left(\frac{\beta}{\beta+1}\right)^{\alpha+x} \\ &= \binom{\alpha+x-1}{x} \left(\frac{1}{\beta+1}\right)^\alpha \left(1 - \frac{1}{\beta+1}\right)^x, \end{aligned}$$

em que x e α são inteiros ($x \geq 0$ e $\alpha > 0$) e $\beta > 0$.

Note que a distribuição marginal de X é Binomial Negativa com $r = \alpha$ e $p = 1/(\beta + 1)$, sendo

$$\begin{aligned} \mathrm{E}[X] &= \mu = \frac{r(1-p)}{p} \\ \mathrm{Var}[X] &= \frac{r(1-p)}{p^2} = \mu + \frac{1}{r}\mu^2. \end{aligned}$$

Se $r \to \infty$ e $p \to 1$ com média μ constante, $P(X = x) \to \frac{e^{-\mu}\mu^x}{x!}$.

O artigo "*Fitting the Negative Binomial Distribution to Biological Data*" [16] traz inúmeras ilustrações de ajuste da Binomial Negativa a dados biológicos com superdispersão.

1.3.2. Poisson Generalizada

Consul e Jain [1] apresentaram a distribuição de Poisson Generalizada (GPD - *Generalized Poisson distribution*) com os parâmetros $\theta > 0$ e $0 \leq \lambda < 1$, sendo

$$P(X = x) = \frac{\theta(\theta + x\lambda)^{x-1}}{x!} e^{-\theta - x\lambda}.$$

A média e a variância de X são dadas por

$$\begin{aligned} \mathrm{E}[X] &= \frac{\theta}{1-\lambda} \\ \mathrm{Var}[X] &= \frac{\theta}{(1-\lambda)^3}. \end{aligned}$$

A depender de λ, a variância dessa distribuição pode ser igual ou maior do que a sua média. Note que, se $\lambda = 0$, $X \sim \mathrm{Poisson}(\theta)$.

1.3.3. Poisson Dupla

Efron [2] propôs a distribuição de Poisson dupla no contexto da família exponencial dupla. A distribuição é obtida como uma combinação exponencial de duas Poisson's, resultando na distribuição Poisson dupla com os parâmetros μ e θ. Seja X uma variável aleatória com distribuição Poisson dupla, sua densidade pode ser escrita na forma

$$P(X = x) = f(x, \mu, \theta) = \theta^{1/2} e^{-\theta\mu} \frac{e^{-x} x^x}{x!} \left(\frac{e\mu}{x}\right)^{\theta x}$$

na qual $x = 0, 1, 2, ...$

A densidade exata da Poisson dupla é dada por

$$P(X = x) = c(\mu, \theta) f(x, \mu, \theta)$$

em que $c(\mu, \theta)$ é a constante de normalização com

$$\frac{1}{c(\mu, \theta)} = \sum_{x=0}^{\infty} f(x, \mu, \theta) \approx 1 + \frac{1 - \theta}{12\mu\theta} \left(1 + \frac{1}{\mu\theta}\right).$$

Essa distribuição tem média e variância aproximadamente iguais a μ e μ/ϕ, respectivamente. Desta forma, essa distribuição permite modelar tanto superdispersão ($\phi < 1$) quanto subdispersão ($\phi > 1$). Quando $\phi = 1$, temos a distribuição de Poisson.

1.3.4. Conway-Maxwell-Poisson

Proposta por Conway e Maxwell [4] e explorada por Shmueli et al. [5], a distribuição de Conway-Maxwell-Poisson é uma generalização da distribuição de Poisson com dois parâmetros e pode ser escrita na forma

$$P(X = x) = \frac{1}{N_{\lambda,\nu}} \frac{\lambda^x}{(x!)^\nu}, \quad x \in \mathbb{Z}^+,$$

em que $\lambda \geq 0$, $\nu \geq 0$ e $N_{\lambda,\nu}$ é uma constante normalizadora definida como

$$N_{\lambda,\nu} = \sum_{i=1}^{\infty} = \frac{\lambda^i}{(i!)^\nu}.$$

A inclusão do parâmetro ν permite que a variância de X seja maior ou menor que a sua média. No caso em que $\nu = 1$, temos a distribuição de Poisson, uma vez que $N_{\lambda,\nu} = e^\lambda$.

Os momentos podem ser obtidos de forma recursiva por

$$\mathrm{E}(X^{r+1}) = \begin{cases} \lambda \mathrm{E}(X+1)^{1-\nu} & r = 0, \\ \lambda \frac{d}{d\lambda}\mathrm{E}(X^r) + \mathrm{E}(X)\mathrm{E}(X^r) & r > 0. \end{cases}$$

1.3.5. Poisson Inflado com Zeros

O modelo de Poisson inflado com zeros (*Zero Inflated Poisson model* - ZIP) é uma modificação da Poisson que permite modelar dados de contagem com muitos zeros entre os valores observados. Seja $X = (x_1, ..., x_n)'$, a ideia básica é que os dados provêm de dois estados. No primeiro, os dados são sempre zero com probabilidade π e no segundo os dados seguem a distribuição de Poisson(λ) com probabilidade $1 - \pi$. As duas componentes são descritas a seguir:

$$X = \begin{cases} 0 & \text{com probabilidade } \pi; \\ \text{Poisson}(\lambda) & \text{com probabilidade } 1 - \pi. \end{cases}$$

Desta forma, a variável X assume os seguintes valores

$$X = \begin{cases} 0 & \text{com probabilidade } \pi + (1-\pi)e^{-\lambda}; \\ x & \text{com probabilidade } (1-\pi)\frac{e^{-\lambda}\lambda_i^x}{x!},\ x = 0, 1, 2, ... \end{cases}$$

A média e a variância de X são $\lambda(1-\pi)$ e $\lambda(1-\pi)(1+\lambda\pi)$, respectivamente.

1.3.6. Nova Poisson-Lindley Generalizada

Proposta por Bhati et al. [6], a distribuição Nova Poisson-Lindley Generalizada – *New Generalized Poisson-Lindley* (NGPL) é obtida de uma Poisson considerando que seu parâmetro λ segue a distribuição Lindley com dois parâmetros (TPLD(θ, α)), sugerida por Shanker et al. [20] e definida como,

$$g(x; \alpha, \theta) = \frac{\theta^2}{(\theta + \alpha)}(1 + \alpha x)e^{-\theta x}, \quad x, \alpha, \theta > 0.$$

Desta forma, uma variável aleatória X é dita ter distribuição NGPL se,

$$X|\lambda \sim \text{Poisson}(\lambda)$$

$$\lambda|\theta, \alpha \sim \text{TPLD}(\theta, \alpha)$$

para $\lambda > 0$ e $\theta, \alpha > 0$.

Assim, a distribuição não-condicional é dada por

$$f(x; \alpha, \theta) = \frac{\theta^2}{(\theta + \alpha)(1 + \theta)^{x+1}}\left(1 + \frac{\alpha(x + 1)}{1 + \theta}\right), \; x = 0, 1, 2, ...$$

com $\alpha, \theta > 0$.

Os dois primeiros momentos são dados por

$$\mu_1' = \frac{2\alpha + \theta}{\theta(\alpha + \theta)} \quad \text{e} \quad \mu_2' = \frac{2\alpha(\theta + 3) + \theta(\theta + 2)}{\theta^2(\alpha + \theta)}.$$

1.3.7. Conway-Maxwell-Poisson Estendida

A Conway-Maxwell-Poisson Estendida – *Extended Conway-Maxwell-Poisson* (ECOMP) é uma extensão com quatro parâmetros da distribuição Conway-Maxwell-

Poisson e foi proposta por Chakraborty e Imoto [7]. Essa distribuição, uma das mais recentes generalizações da Poisson, é definida por

$$P(X=x)=\frac{\{(\nu)_x\}^\beta}{{}_1S^\beta_{a-1}(\nu;1;p)}\frac{p^x}{(x!)^\alpha}=\frac{\{\Gamma(\nu+x)\}^\beta}{(\Gamma\nu)^\beta\,{}_1S^\beta_{a-1}(\nu;1;p)}\frac{p^x}{(x!)^\alpha}$$

e o espaço paramétrico é dado por

$$\Theta_{ECOMP}=\{\nu\geq 0,p>0,\alpha>\beta\}\cup\{\nu>0,0<p<1,\alpha=\beta\}.$$

A constante de normalização ${}_1S^\beta_{\alpha-1}(\nu;1;p)$ é oriunda da série tipo hipergeométrica

$${}_mS^\beta_a(a_1,a_2,...,a_m;b;p)=\sum_{k=0}^{\infty}\frac{\{(a_1)_k\}^\beta(a_2)_k...(a_m)_k}{\{(b)_k\}^\alpha}\frac{p^k}{k!},$$

em que $(a)_k=a(a+1)...(a+k-1)=\Gamma(a+k)/\Gamma a$.

Os momentos dessa distribuição podem ser obtidos pela função

$$\mathrm{E}(X^{[r]})=\mu^{[r]}=\frac{\{(\nu)_r\}^\beta p^r}{(r!)^{\alpha-1}}\frac{{}_1S^\beta_{a-1}(\nu+r;r+1;p)}{{}_1S^\beta_{a-1}(\nu;1;p)}.$$

Capítulo 2

Distribuição Touchard

Desde o início do século XX, diversas generalizações da Poisson surgiram com o objetivo de adequar a clássica distribuição de Poisson a dados com subdispersão, superdispersão e excesso de zeros. Muitas dessas propostas contornam o problema da superdispersão, outras se voltaram para o excesso de zeros nos dados. Em alguns casos, suas funções de probabilidade são bastante complexas.

Neste capítulo será apresentada a distribuição Touchard proposta por Matsushita et al. [9]. Uma generalização da Poisson que, com a adição de mais um parâmetro, permite modelar dados com subdispersão, superdispersão e excesso de zeros.

O nome da distribuição está relacionado aos polinômios de Touchard [21], já utilizados em formulações de modelos aplicados em problemas de passeios aleatórios [22] e sistema de filas [23].

2.1. Definição

Seja X uma variável aleatória inteira não-negativa, $k \in \mathbb{N}$, com distribuição de probabilidade definida como,

$$p_k = P[X = k] = \frac{\lambda^k (k+1)^\delta}{k! \tau(\lambda, \delta)}, \tag{2.1}$$

em que $\lambda > 0$ e $\delta \in \mathbb{R}$ são parâmetros da distribuição e a função

$$\tau(\lambda, \delta) = \sum \frac{\lambda^j (j+1)^\delta}{j!} \tag{2.2}$$

normaliza (2.1) e está relacionada aos polinômios de Touchard e ao momento de ordem δ de uma Poisson deslocada. Desta forma, $X \sim \text{Touchard}(\lambda, \delta)$ definida em (2.1) é uma generalização da Poisson e, para $\delta = 0$, $X \sim \text{Poisson}(\lambda)$.

A Figura 2.1 ilustra como a distribuição Touchard pode ser flexível, permitindo modelar diferentes formas de distribuição.

2.2. Momentos

O r-ésimo momento de uma variável com distribuição Touchard é dado pela série

$$\text{E}[X^r] = \sum_{j=0}^{r} \binom{r}{j} \frac{(-1)^{r-j} \tau(\lambda, \delta + j)}{\tau(\lambda, \delta)}, \tag{2.3}$$

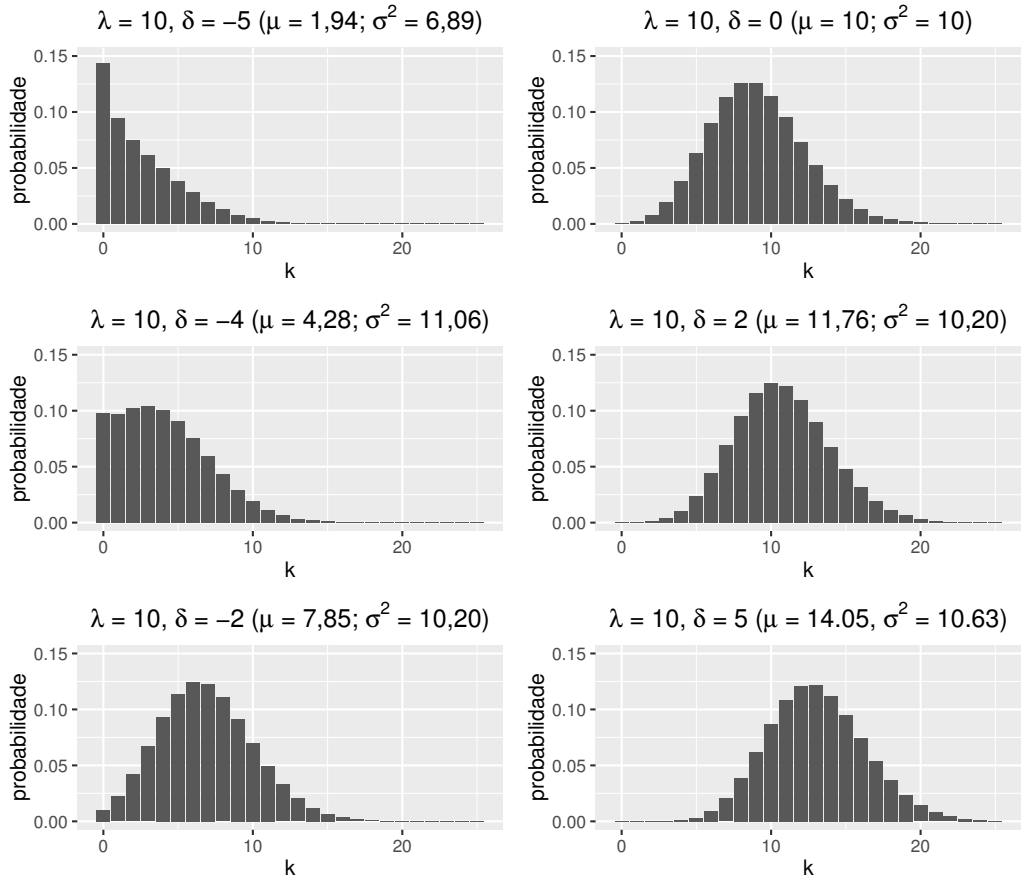

Figura 2.1: Exemplos da distribuição Touchard com $\lambda = 10$ e δ variando entre -5 e 5.

e sua função geratriz de momentos é

$$M_X(q) = \mathrm{E}\left[e^{qX}\right] = \frac{\tau(\lambda e^q, \delta)}{\tau(\lambda, \delta)},$$

em que $q \in \mathbb{R}$.

A média de X pode ser expressa por

$$\begin{aligned} \mu = \mathrm{E}[X] &= \frac{\tau(\lambda, \delta+1)}{\tau(\lambda,\delta)} - 1 \\ &= \lambda.\mathrm{E}\left[\left(\frac{X+2}{X+1}\right)^{\delta}\right], \end{aligned} \tag{2.4}$$

e a variância por

$$\begin{aligned} \sigma^2 = \mathrm{Var}[X] &= \frac{\tau(\lambda, \delta+2)}{\tau(\lambda,\delta)} - \left[\frac{\tau(\lambda, \delta+1)}{\tau(\lambda,\delta)}\right]^2 \\ &= \lambda\mathrm{E}\left[(X+1)\left(\frac{X+2}{X+1}\right)^{\delta}\right] - \mu^2. \end{aligned} \tag{2.5}$$

Note que $\mu > \lambda$ se $\delta > 0$ e $\mu < \lambda$ se $\delta < 0$. Pode-se avaliar a relação entre a média e a variância por meio da razão $r = \sigma^2/\mu$. Verifica-se que

$$r = \frac{\mathrm{E}\left[(X+1)\left(\frac{X+2}{X+1}\right)^{\delta}\right]}{\mathrm{E}\left[\left(\frac{X+2}{X+1}\right)^{\delta}\right]} - \mu.$$

Quando $\delta = 0$, temos a distribuição de Poisson e, naturalmente, $r = 1$. Para $\delta > 0$, temos $r < 1$ (subdispersão) e nos casos em que $\delta < 0$, então $r > 1$, indicando superdispersão.

2.3. Estatísticas suficientes

A distribuição Touchard faz parte da família exponencial pois sua distribuição de probabilidade pode ser escrita na forma

$$p_k = P[X = k] = \frac{1}{k!} \exp\{k \ln(\lambda) + \delta \ln(k+1) - \ln[\tau(\lambda, \delta)]\}. \tag{2.6}$$

Seja $x_1, ..., x_n$ uma amostra aleatória de n realizações da distribuição Touchard, a função de verossimilhança pode ser escrita como

$$L(\lambda, \delta \mid \{x_i\}) = \left(\prod_i x_i!\right)^{-1} \lambda^{S_1} e^{\delta S_2} [\tau(\lambda, \delta)]^{-n}, \tag{2.7}$$

em que $S_1 = \sum_i x_i$ e $S_2 = \sum_i \ln(x_i + 1)$ são estatísticas suficientes, de acordo com o Teorema da Fatoração.

2.4. Estimadores de máxima verossimilhança

Seja $l(\lambda, \delta) = \ln L(\lambda, \delta \mid \{x_i\})$, a maximização de $l(\lambda, \delta)$ pode ser feita por meio das duas primeiras derivadas de $\tau(\lambda, \delta)$ em relação aos parâmetros λ e δ, dadas por

$$\begin{aligned} \frac{\partial \tau(\lambda, \delta)}{\partial \lambda} &= \frac{\tau(\lambda, \delta)}{\lambda} \mu, \\ \frac{\partial \tau(\lambda, \delta)}{\partial \delta} &= \tau(\lambda, \delta) \mathrm{E}\{\ln[X + 1]\}. \end{aligned}$$

e

$$\frac{\partial^2\tau(\lambda,\delta)}{\partial\lambda^2} = \tau(\lambda,\delta)\frac{\mathrm{E}[X^2]-\mu}{\lambda^2},$$
$$\frac{\partial^2\tau(\lambda,\delta)}{\partial\delta^2} = \tau(\lambda,\delta)\mathrm{E}\{\ln^2[X+1]\},$$
$$\frac{\partial^2\tau(\lambda,\delta)}{\partial\delta\partial\lambda} = \frac{\tau(\lambda,\delta)}{\lambda}\mathrm{E}\{X\ln(X+1)\}.$$

Com isso, as equações de máxima verossimilhança são

$$\begin{cases} S_1 - n\mu = 0; \\ S_2 - n\mathrm{E}\left\{\ln[X+1]\right\} = 0. \end{cases} \tag{2.8}$$

As estimativas de máxima verossimilhança para λ e δ podem ser encontradas através do método de Newton. A matriz hessiana é dada por

$$H = \begin{pmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{pmatrix},$$

em que,

$$H_{11} = \frac{\partial^2 l(\lambda,\delta)}{\partial\lambda^2} = \frac{n[\sigma^2+(\bar{x}-\mu)]}{\lambda^2};$$
$$H_{22} = \frac{\partial^2 l(\lambda,\delta)}{\partial\delta^2} = -n\mathrm{Var}[\ln(X+1)];$$
$$H_{12} = H_{21} = \frac{\partial^2 l(\lambda,\delta)}{\partial\lambda\partial\delta} = -\frac{n\mathrm{Cov}[X,\ln(X+1)]}{\lambda}.$$

2.5. Fórmulas recursivas

Nesta seção são apresentadas equações úteis para a implementação de algoritmos envolvendo a distribuição Touchard. Considere X uma variável aleatória com distribuição Touchard(λ, δ).

2.5.1. Função de distribuição

A função de distribuição, definida pela equação (2.1), pode ser obtida na forma recursiva

$$p_{k+1} = \frac{\lambda}{k+1}\left(\frac{k+2}{k+1}\right)^{\delta} p_k. \tag{2.9}$$

2.5.2. Função $\tau(\lambda, \delta)$

A constante de normalização $\tau(\lambda, \delta)$, definida pela função (2.2), pode ser calculada numericamente utilizando a forma alternativa

$$\tau(\lambda, \delta) = \sum_{j \in \mathbb{N}} A_j, \tag{2.10}$$

em que

$$A_{j+1} = \frac{\lambda}{j+1}\left(\frac{j+2}{j+1}\right)^{\delta} A_j,$$

com $A_0 = 1$. Quando $\delta = 0$, essa aproximação não é necessária, uma vez que

$\tau(\lambda, 0) = e^{\lambda}$.

2.5.3. Momentos de X

De forma similar, os momentos de X definidos pela equação (2.3) podem ser obtidos pela fórmula

$$\mathrm{E}\left[X^r\right] = \frac{1}{\tau(\lambda, \delta)} \sum_{j=1}^{\infty} A_{r,j}, \tag{2.11}$$

para $r \geq 1$, em que $A_{r,1} = \lambda 2^{\delta}$ e

$$A_{r,j+1} = \frac{\lambda}{j+1} \left(\frac{j+1}{j}\right)^r \left(\frac{j+2}{j+1}\right)^{\delta} A_{r,j}.$$

2.5.4. Momentos de $\ln(X+1)$

O r-ésimo momento de $\ln(X+1)$, com $r \geq 1$, pode ser escrito na forma

$$\mathrm{E}\left[\ln^r(X+1)\right] = \frac{1}{\tau(\lambda, \delta)} \sum_{j=1}^{\infty} B_{r,j}, \tag{2.12}$$

em que $B_{r,1} = \lambda 2^{\delta} (\ln 2)^r$ e

$$B_{r,j+1} = \frac{\lambda}{j+1} \left(\frac{\ln(j+2)}{\ln(j+1)}\right)^r \left(\frac{j+2}{j+1}\right)^{\delta} B_{r,j}.$$

2.5.5. Valor Esperado de $X\ln(X+1)$

A esperança de $g(X) = X\ln(X+1)$ pode ser obtida por

$$\mathrm{E}\left[X\ln(X+1)\right] = \frac{1}{\tau(\lambda,\delta)}\sum_{j=1}^{\infty} C_j, \tag{2.13}$$

em que $C_1 = \lambda 2^{\delta}\ln 2$ e

$$C_{j+1} = \frac{\lambda\ln(j+2)}{\lambda\ln(j+1)}\left(\frac{j+2}{j+1}\right)^{\delta} C_j.$$

Capítulo 3

Regressão Touchard

Neste capítulo será introduzida a regressão Tourchard, proposta como alternativa de análise de regressão para modelagem de dados de contagem em que a distribuição de Poisson não é adequada.

A regressão Touchard assume que a variável resposta Y tem distribuição Touchard e que seus parâmetros λ e δ podem ser modelados por covariáveis. Ou seja, na regressão Touchard, as variáveis explicativas podem ser utilizadas para explicar seu parâmetro de posição (λ) ou ainda seu parâmetro de dispersão (δ).

Em geral, é mais intuitivo a modelagem do parâmetro λ, contudo, podem ocorrer situações em que certa covariável está mais relacionada com a dispersão dos dados do que com a sua média.

3.1. O modelo

Seja $y = (y_1, ..., y_n)'$ uma amostra aleatória em que y_i vem de uma distribuição Touchard (2.1) com parâmetros λ_i e δ_i, o modelo de regressão Touchard pode ser

obtido assumindo que

$$\lambda_i = e^{x_i'\beta} \tag{3.1}$$

e

$$\delta_i = z_i'\alpha \tag{3.2}$$

em que $x_i' = (x_{i1}, ..., x_{ip})$ e $z_i' = (z_{i1}, ..., z_{iq})$ são observações de covariáveis fixas e conhecidas e $\beta = (\beta_1, ..., \beta_p)' \in \mathbb{R}^p$ e $\alpha = (\alpha_1, ..., \alpha_q)' \in \mathbb{R}^q$ são vetores de parâmetros desconhecidos.

3.2. Estimadores de máxima verossimilhança

A função de log-verossimilhança baseada na amostra de n observações independentes é

$$l(\beta, \alpha) = \sum_{i=1}^{n} \ln p_{y_i}, \tag{3.3}$$

em que

$$\ln p_{y_i} = y_i \ln \lambda_i + \delta_i \ln(y_i + 1) - \ln y_i! - \ln \tau(\lambda_i, \delta_i), \tag{3.4}$$

com λ_i e δ_i funções de β e α, respectivamente, conforme definido em (3.1) e (3.2).

Sejam $y^* = (\ln(y_1+1), ..., \ln(y_n+1))'$, $\mu = (\mu_1, ..., \mu_n)'$ e $\mu^* = (\mu_1^*, ..., \mu_n^*)'$, em que $\mu_i^* = \mathrm{E}\left[\ln(Y_i+1)\right]$. A função escore, obtida pela derivada de (3.3) em relação aos coeficientes β e α é dada por

$$\begin{cases} U_\beta = X'(y-\mu) \\ U_\alpha = Z'(y^*-\mu^*) \end{cases} \tag{3.5}$$

com X e Z sendo as matrizes de delineamento nas quais suas i-ésimas linhas são x_i' e z_i', respectivamente.

3.3. Matriz Hessiana

Sejam $\sigma_i^{*2} = \mathrm{Var}[\ln(Y_i+1)]$ e $\gamma_i = \mathrm{Cov}[Y_i, \ln Y_i+1]$, a matriz Hessiana é definida por

$$H = \begin{pmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{pmatrix}, \tag{3.6}$$

com

$$\begin{aligned} H_{11} &= -X'DX; \\ H_{22} &= -Z'D^*Z; \\ H_{12} &= H_{21}' = -X'CZ. \end{aligned}$$

em que $D = diag(\sigma_i^2)$, $D^* = diag(\sigma_i^{*2})$ e $C = diag(\gamma_i)$. Considerando que H é uma matriz particionada, sua inversa pode ser obtida conforme abaixo,

$$H^{-1} = \begin{pmatrix} H^{11} & H^{12} \\ H^{21} & H^{22} \end{pmatrix}, \tag{3.7}$$

com

$$\begin{aligned} H^{11} &= (H_{11} - H_{12}H_{22}^{-1}H_{21})^{-1}; \\ H^{22} &= (H_{22} - H_{21}H_{11}^{-1}H_{12})^{-1}; \\ H^{12} &= H^{21'} = -H^{11}H_{12}H_{22}. \end{aligned}$$

Capítulo 4

Implementação computacional

Desenvolvemos um pacote na linguagem R disponibilizando as função de probabilidade (`dtouchard`), função de distribuição (`ptouchard`), função quantil (`qtouchard`) e um gerador de números aleatórios (`rtouchard`) de uma distribuição Touchard com parâmetros λ e δ.

Essas funções seguem o mesmo padrão das clássicas funções disponíveis no R, como, por exemplo, as funções das distribuições Normal (`dnorm`, `pnorm`, `qnorm` e `rnorm`) e Poisson (`dpois`, `ppois`, `qpois` e `rpois`), tanto na entrada dos dados quanto em relação às saídas (*output*).

Também foi disponibilizada no pacote Touchard a função $\tau(\lambda, \delta)$ que normaliza a função de distribuição da Touchard, como visto na Equação (2.1). A função `moment.touch` retorna o r-ésimo momento da distribuição Touchard, enquanto que a função `mle.touch` devolve, dada uma amostra $(x_1, x_2, \ldots, x_n)$, os estimadores de máxima verossimilhança da distribuição Touchard. A regressão Touchard, apresentada no Capítulo 3, também pode ser ajustada por meio desse pacote.

As estimativas de máxima verossimilhança para os parâmetros λ e δ da distribuição e para os parâmetros da regressão ($\beta_0, \beta_1, \ldots, \beta_k$ e α_0) foram obtidas por

meio das funções escores e do algoritmo de Newton. Uma outra versão das funções `mle.touch` e `touch.reg` utilizou a função de otimização interna do R, `nlminb` [1].

O pacote `Touchard Distribution` pode ser baixado no endereço eletrônico `https://goo.gl/SsrPk6`. Como trabalho futuro, serão implementadas nesse pacote a distribuição e a regressão Touchard com três parâmetros, bem como a disponibilização deste pacote no CRAN (*The Comprehensive R Archive Network*).

[1] https://stat.ethz.ch/R-manual/R-devel/library/stats/html/nlminb.html

Capítulo 5

Distribuição Touchard com três parâmetros

O excesso de zeros, como visto neste trabalho, torna inadequada a modelagem baseada na distribuição de Poisson. A distribuição Touchard pode ser utilizada como uma alternativa para modelar esses tipos de dados.

O excesso de zeros pode estar associado, não apenas a eventos raros, mas a dados ausentes (*missing values*). Pode ocorrer ainda, por exemplo, nas situações em que a ocorrência do evento é rara, mas uma vez ocorrido, tem-se um grande número de falhas no produto [8]. Nesses casos, podemos ter uma distribuição inflada com zeros, ou seja, uma alta concentração de observações no ponto zero e uma outra massa de dados concentrada mais à direita, com média condicional muito maior que zero.

Nesse sentido, Matsushita et al. [24] propuseram a Touchard com três parâmetros, Touchard$(\lambda, \delta, \theta)$. A inclusão do parâmetro θ, chamado parâmetro de deslocamento, permite modelar dados com essas características.

5.1. Definição

Seja X uma variável aleatória inteira não-negativa com k pertencente ao conjunto dos números naturais $\mathbb{N}$, dizemos que X tem distribuição Touchard$(\lambda, \delta, \theta)$ se

$$p_k = P[X = k] = \frac{\lambda^k (k+\theta)^\delta}{k!\tau(\lambda, \delta, \theta)}, \tag{5.1}$$

em que $\lambda > 0$, $\delta \in \mathbb{R}$ e $\theta > 0$ são parâmetros da distribuição e a função

$$\tau(\lambda, \delta, \theta) = \sum \frac{\lambda^j (j+\theta)^\delta}{j!} \tag{5.2}$$

normaliza (5.1). É fácil notar que, se $\delta = 0$, $X \sim$ Poisson(λ).

Recursivamente, para $p_k > 0$, podemos escrever (5.1) como

$$p_{k+1} = \frac{\lambda}{k+1}\left(1 + \frac{1}{k+\theta}\right)^\delta p_k. \tag{5.3}$$

A Figura 5.1 traz exemplos de formas das distribuição Touchard com três parâmetros de acordo com as variações dos seus parâmetros.

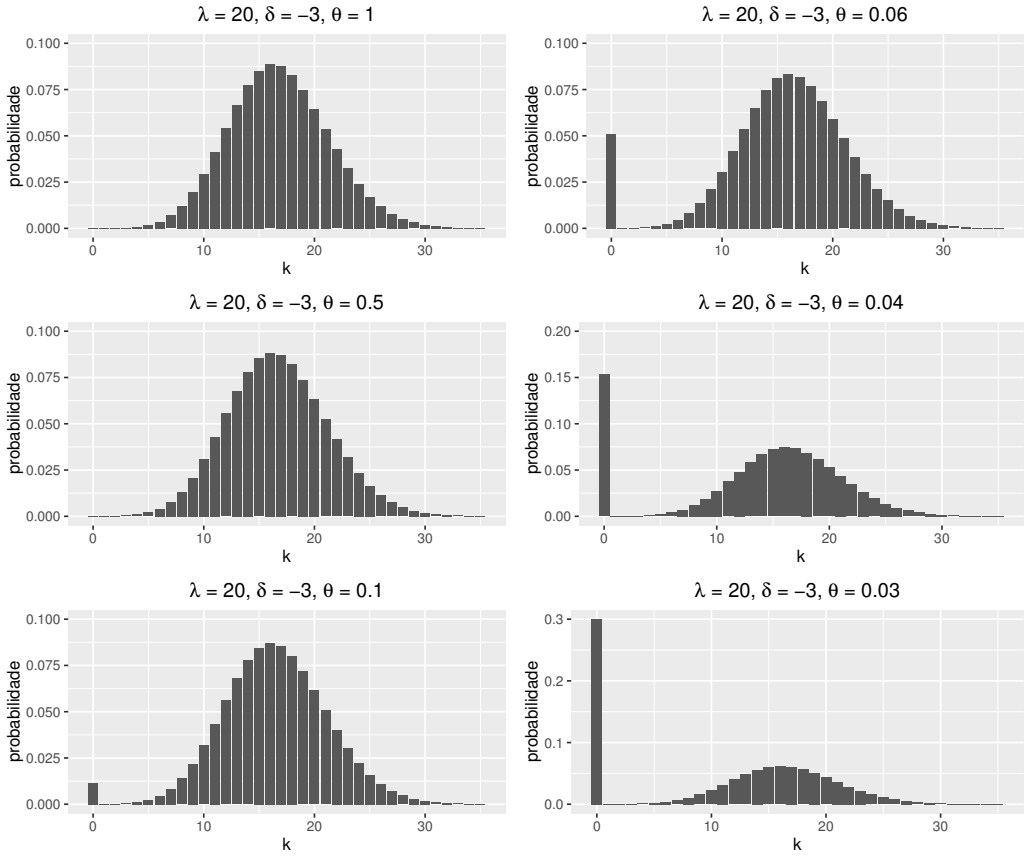

Figura 5.1: Exemplos da distribuição Touchard com $\lambda = 10$, $\delta = -3$ e θ variando entre 0,03 e 1,00.

5.2. Momentos

O r-ésimo momento de uma variável com distribuição Touchard com três parâmetros é obtido a partir da série

$$\mathrm{E}[X^r] = \sum_{j=0}^{r} \binom{r}{j} \frac{(-\theta)^{r-j}\tau(\lambda, \delta + j, \theta)}{\tau(\lambda, \delta, \theta)}, \tag{5.4}$$

e sua função geratriz de momentos é

$$M_X(q) = \mathrm{E}\left[e^{qX}\right] = \frac{\tau(\lambda e^q, \delta, \theta)}{\tau(\lambda, \delta, \theta)},$$

em que $q \in \mathbb{R}$.

Assim, com base na equação (5.4), obtém-se a média

$$\begin{aligned} \mu = \mathrm{E}[X] &= \frac{\tau(\lambda, \delta+1, \theta)}{\tau(\lambda, \delta, \theta)} - \theta \\ &= \lambda \mathrm{E}\left[\left(1 + \frac{1}{X+\theta}\right)^{\delta}\right], \end{aligned} \tag{5.5}$$

e a variância

$$\begin{aligned} \sigma^2 = \mathrm{Var}[X] &= \frac{\tau(\lambda, \delta+2, \theta)}{\tau(\lambda, \delta, \theta)} - \left[\frac{\tau(\lambda, \delta+1, \theta)}{\tau(\lambda, \delta, \theta)}\right]^2 \\ &= \lambda \mathrm{E}\left[(X+1)\left(1 + \frac{1}{X+\theta}\right)^{\delta}\right] - \mu^2. \end{aligned} \tag{5.6}$$

Seja $\theta > 0$, quando $\mu > \lambda$, tem-se $\delta > 0$. Do contrário ($\mu < \lambda$), $\delta < 0$. A relação entre a média e a variância pode ser avaliada por meio da razão $r = \sigma^2/\mu$, na qual

$$r = \frac{\mathrm{E}\left[(X+1)\left(1 + \frac{1}{X+\theta}\right)^{\delta}\right]}{\mathrm{E}\left[\left(1 + \frac{1}{X+\theta}\right)^{\delta}\right]} - \mu.$$

Quando a distribuição for Poisson ($\delta = 0$), $r = 1$. Para $\delta > 0$, tem-se $r < 1$ (subdispersão), nos casos em que $\delta < 0$, então $r > 1$ (superdispersão).

5.3. Estimadores de máxima verossimilhança

Considere $x_1, x_2, \ldots, x_n$ um conjunto de n observações independentes da distribuição Touchard$(\lambda, \delta, \theta)$, a função log-verossimilhança pode ser escrita como

$$l(\lambda, \delta, \theta | x_i) = -\sum_{i=1}^{n} \ln x_i! + \ln\lambda \sum_{i=1}^{n} x_i + \delta \sum_{i=1}^{n} \ln(x_i + \theta) - n\ln\tau(\lambda, \delta, \theta) \tag{5.7}$$

e as equações para maximização da verossimilhança são

$$\begin{cases} \sum_{i=1}^{n} x_i - n\mu = 0, \\ \sum_{i=1}^{n} \ln(x_i + \theta) - n\mathrm{E}[\ln(X + \theta)] = 0, \\ \sum_{i=1}^{n} (x_i + \theta)^{-1} - n\mathrm{E}[(x_i + \theta)^{-1}] = 0. \end{cases} \tag{5.8}$$

Os estimadores de momentos de λ, δ e θ que satisfazem (5.7) coincidem com seus respectivos estimadores de máxima verossimilhança $(\hat{\lambda}, \hat{\delta}, \hat{\theta})$.

A matriz hessiana no ponto $(\hat{\lambda}, \hat{\delta}, \hat{\theta})$ é

$$\hat{H} = \begin{pmatrix} \hat{H}_{11} & \hat{H}_{12} & \hat{H}_{13} \\ \hat{H}_{12} & \hat{H}_{22} & \hat{H}_{23} \\ \hat{H}_{13} & \hat{H}_{23} & \hat{H}_{33} \end{pmatrix}, \tag{5.9}$$

sendo

$$
\begin{aligned}
H_{11} &= -n\hat{\sigma}^2/\hat{\lambda}^2, \\
H_{22} &= -n\mathrm{Var}[\ln(X+\hat{\theta})], \\
H_{12} &= -n\mathrm{Cov}[X, \ln(X+\hat{\theta})]/\hat{\lambda}, \\
H_{13} &= -n\hat{\delta}\mathrm{Cov}[X, (X+\hat{\theta})^{-1}]/\hat{\lambda}, \\
H_{23} &= -n\hat{\delta}\mathrm{Cov}[\ln(X+\hat{\theta}), (X+\hat{\theta})^{-1}], \\
H_{33} &= -n\hat{\delta}^2\mathrm{var}[(X+\hat{\theta})^{-1}] - n\hat{\delta}\left\{\sum_{i=1}^{n}(x_i+\hat{\theta})^{-2}/n - \mathrm{E}[(X+\hat{\theta})^{-2}]\right\}.
\end{aligned}
$$

Para se obter valores numéricos, a equação (5.2) deve ser truncada (exceto para $\delta = 0$, pois $\tau(\lambda, 0, \theta) = \mathrm{e}^{\lambda}$), sendo sugerida a forma recursiva a seguir, evitando-se o termo fatorial,

$$
\tau(\lambda, \delta, \theta) = \sum_{j=1}^{m} A_j, \tag{5.10}
$$

na qual $A_{j+1} = \frac{\lambda}{j+1}\left(\frac{j+1+\theta}{j+\theta}\right)^{\delta} A_j$, com $A_0 = \theta^{\delta}$, e m depende da precisão numérica predefinida. Assim, os momentos de X necessários para o cálculo da matriz hessiana podem ser calculados diretamente das equações (5.4) e (5.10). Desta forma, o r-ésimo momento de $\ln(X+\theta)$, $r \geq 1$, pode ser expresso por

$$
\mathrm{E}[\ln^r(X+\theta)] = \frac{\eta(r, \lambda, \delta, \theta)}{\tau(\lambda, \delta, \theta)},
$$

em que

$$\eta(r,\lambda,\delta,\theta)=\sum_{j=0}^{\infty}B_{r,j}, \tag{5.11}$$

e $B_{r,j+1}=\frac{\lambda}{j+1}\left(\frac{\ln(j+1+\theta)}{\ln(j+\theta)}\right)^{r}\left(\frac{j+1+\theta}{j+\theta}\right)^{\delta}B_{r,j}$. Para $\theta\neq 1$ a forma recursiva inicia com $B_{r,0}=\theta^{\delta}(\ln\theta)^{r}$, para $\theta\neq 1$. Mas se $\theta=1$, $B_{r,1}=\lambda(1+\theta)^{\delta}[\ln(1+\theta)]^{r}$ e $B_{r,0}=0$.

Baseado em (5.2) e (5.11), os momentos cruzados vistos em (5.9) podem ser escritos como

$$\begin{aligned}
\mathrm{E}[X\ln(X+\theta)] &= \lambda\frac{\eta(1,\lambda,\delta,\theta+1)}{\tau(\lambda,\delta,\theta)};\\
\mathrm{E}[\ln(X+\theta)(X+\theta)^{-1}] &= \frac{\eta(1,\lambda,\delta-1,\theta)}{\tau(\lambda,\delta,\theta)};\\
\mathrm{E}[X(X+\theta)^{-1}] &= \lambda\frac{\tau(\lambda,\delta-1,\theta+1)}{\tau(\lambda,\delta,\theta)}.
\end{aligned}$$

Capítulo 6

Aplicações

Neste capítulo será apresentada uma série de aplicações da distribuição Touchard comparando o seu desempenho com as distribuições propostas nos trabalhos originais. Também serão utilizados dados com relevância prática com o objetivo de ilustrar situações em que a distribuição Touchard é preferível à clássica Poisson.

Para verificar o ajuste dos modelos aos dados foi avaliada a discrepância entre os dados observados e os valores esperados. Essa discrepância foi medida por meio do teste Qui-quadrado de bondade do ajuste (χ^2),

$$\chi^2_{k-p-1} = \sum_{i=0}^{k} \frac{(Observado_i - Esperado_i)^2}{Esperado_i},$$

em que k é o número de categorias ou classes (após combinação das classes) e p é o número de parâmetros estimados a partir dos dados.

As classes com valores esperados muito pequenos foram agrupadas de forma que, preferivelmente, não tivessem valores esperados inferiores a 5. O p-valor foi obtido assumindo que a estatística χ^2 tem $(k - p - 1)$ graus de liberdade (GL).

Também foram utilizados os critérios de informação de Akaike (AIC) [25] e bayesiano (BIC) para a comparação entres os diferentes modelos propostos.

6.1. Dados biológicos: Touchard x Binomial Negativa

Frequentemente em dados biológicos tem-se a variância claramente maior que sua média, caracterizando a superdispersão. Bliss e Fisher [16] mostraram essa característica e propuseram o ajuste desses dados utilizando a distribuição Binomial Negativa.

Nesta seção serão utilizadas algumas da aplicações utilizadas no artigo de Bliss e Fisher [16] para ilustrar o desempenho da distribuição Touchard frente a dados biológicos com superdispersão. As tabelas a seguir reproduzem os resultados encontrados naquele artigo em comparação com as estimativas decorrentes do ajuste pela generalização da Poisson proposta nesta dissertação.

A Tabela 6.1.1 traz dados sobre a contagem de ácaros vermelhos em folhas de macieira. De acordo com teste χ^2, o ajuste da Touchard com dois parâmetros foi tão bom quanto a modelagem com a Binomial Negativa.

Tabela 6.1.1: Distribuição da contagem de ácaros vermelhos em folhas de macieira.

	Ácaros por folha							
	0	1	2	3	4	5+	χ^2	p-valor
Observado	70	38	17	10	9	6	—	—
Bin. Neg.	69,5	37,6	20,1	10,7	5,7	6,4	2,48	0,478
Touchard	71,1	33,6	20,4	12,1	6,7	6,0	2,32	0,508
Touchard3	70,5	35,2	20,1	11,6	6,4	6,3	1,99	0,370

A distribuição do número de células de levedura por quadrado em um hemocitómetro (Tabela 6.1.2) também foi bem ajustada pelas variações da distribuição Touchard.

Tabela 6.1.2: Distribuição da contagem de células de levedura por quadrado num hemocitómetro.

	Células de levedura								
	0	1	2	3	4	5	6+	χ^2	p-valor
Observado	213	128	37	18	3	1	0	—	—
Bin. Neg.	214,2	122,8	45,0	13,4	3,5	0,9	0,2	3,51	0,476
Touchard	214,7	120,6	46,8	13,8	3,3	0,7	0,1	4,05	0,390
Touchard3	214,2	122,7	45,0	13,4	3,5	0,9	0,2	3,42	0,312

Outro exemplo em que a Touchard se mostrou competitiva em comparação com a distribuição Binomial Negativa se refere a dados sobre o número de acidentes sofridos por mecânicos no período de três meses (Tabela 6.1.3).

Tabela 6.1.3: Distribuição do número de acidentes sofridos por mecânicos no período de três meses.

	Acidentes por mecânico								
	0	1	2	3	4	5	6+	χ^2	p-valor
Observado	296	74	26	8	4	4	2	—	—
Bin. Neg.	296,7	71,0	26,4	11,0	4,8	2,2	1,9	2,56	0,633
Touchard	298,6	67,3	26,9	12,1	5,4	2,3	1,5	3,89	0,421
Touchard3	296,2	73,7	24,7	10,1	4,6	2,3	2,4	1,97	0,579

O ajuste do número de *Liatris aspera* por praça pela Touchard com três parâmetros foi quase perfeito ($\chi^2 = 0,95$; $p - valor = 0,812$), superando, segundo o teste Qui-quadrado de bondade do ajuste, a distribuição Binomial Negativa (Tabela 6.1.4).

Tabela 6.1.4: Distribuição do número de *Liatris aspera* (planta).

	Número de *Liatris aspera*								
	0	1	2	3	4	5	6+	χ^2	p-valor
Observado	7.403	183	34	14	4	1	1	—	—
Bin. Neg.	7.403,1	179,8	40,0	11,5	3,7	1,3	0,6	1,86	0,761
Touchard	7.399,7	192,1	31,3	9,5	3,9	1,8	1,8	3,44	0,486
Touchard3	7.403,1	181,9	37,0	11,6	4,1	1,5	0,8	0,95	0,812

6.2. Exemplos de Consul e Jain

No trabalho publicado por Consul e Jain [1], a Poisson Generalizada – *Generalized Poisson* (GP) com dois parâmetros foi comparada às distribuições Poisson e Binomial Negativa. Com os mesmos exemplos utilizados naquele artigo, o desempenho da Touchard foi avaliado comparando-se as estatísticas Qui-quadrado de bondade do ajuste dos modelos GP, Poisson e Binomial Negativa. Assim como na seção anterior, classes com valores esperados inferiores a 5 foram agrupadas para o cálculo da estatística de teste χ^2.

A ilustração apresentada na Tabela 6.2.1 utiliza os clássicos dados de Bortkiewicz (1898) sobre o número de mortes no exército da Prússia causadas por coices de cavalos ou mulas. Nessa aplicação, podemos verificar que o desempenho da Touchard é idêntico ao da GP.

Tabela 6.2.1: Distribuição do número de mortes por coices de cavalos no exército da Prússia.

	Mortes						
	0	1	2	3	4+	χ^2	p-valor
Observado	109	65	22	3	1	—	—
Poisson	108,7	66,3	20,2	4,1	0,7	0,32	0,851
GP	108,7	66,2	20,2	4,1	0,7	0,33	0,568
Touchard	108,8	66,2	20,2	4,1	0,7	0,33	0,568

A Tabela 6.2.2 apresenta os dados publicados em Greenwood e Yule [14] sobre o número de acidentes sofridos por funcionárias da H. E. Shells em cinco semanas. Consul e Jain utilizaram esses dados em seu trabalho para ilustrar sua generalização da Poisson. Ajustando esses mesmos dados com a Touchard verifica-se um melhor ajuste desse modelo aos dados. Nesse exemplo, o ganho com o acréscimo do terceiro parâmetro da Touchard é mínimo, sendo suficiente o ajuste da Touchard com dois parâmetros.

Tabela 6.2.2: Distribuição do número de acidentes sofridos por funcionárias da H. E. Shells em cinco semanas.

	Acidentes							
	0	1	2	3	4	5+	χ^2	p-valor
Observado	447	132	42	21	3	2	—	—
Bin. Neg.	442,9	138,6	44,4	14,3	4,6	2,2	4,09	0,129
GP	441,8	140,8	43,6	13,9	4,6	2,4	4,83	0,089
Touchard	446,9	131,0	46,5	15,8	4,9	1,8	2,61	0,272
Touchard3	447,2	130,0	47,2	16,0	4,9	1,7	2,54	0,111

Os dados utilizados no exemplo a seguir foram retirados de uma série de aplicações publicadas por Thorndike [26]. A Tabela 6.2.3 traz a comparação entre as distribuições Binomial Negativa, GP e Touchard. De acordo com o teste Qui-quadrado, a Touchard apresentou um excelente ajuste ao dados, com a discrepância entre os valores esperados e observados muito próxima de zero.

Tabela 6.2.3: Distribuição do número de artigos perdidos encontrados no Edifício *Telephone and Telegraph, New York City.*

	Número de artigos									
	0	1	2	3	4	5	6	7+	χ^2	p-valor
Observado	169	134	74	32	11	2	0	1	—	—
Bin. Neg.	166,0	140,4	72,4	29,3	10,3	3,3	1,0	0,4	0,67	0,713
GP	165,8	140,9	72,3	29,2	10,2	3,3	1,0	0,4	0,76	0,683
Touchard	168,9	134,4	74,1	31,1	10,6	3,0	0,7	0,2	0,04	0,979
Touchard3	168,9	134,3	74,2	31,2	10,6	3,0	0,7	0,2	0,04	0,845

6.3. Ajustando dados com excesso de zeros

A alta incidência de zeros, superior ao esperado para uma distribuição Poisson, é uma das causas do fenômeno superdispersão. No contexto de modelos de regressão, um dos pioneiros na modelagem de dados de contagens com excesso de zeros foi Lambert [8] em aplicação na área de defeitos em equipamentos manufaturados.

Esses modelos são conhecidos como *Zero-inflated Poisson* (ZIP).

Para ilustrar o desempenho da regressão Touchard em dados com excesso de zeros, foi utilizada uma aplicação explorada por Ridout et al. [27] em sua revisão sobre os modelos inflados com zero. Os dados ajustados tratam do número de raízes produzidas por 270 brotos no cultivo de maçã. Esses brotos foram dispostos em condições experimentais de concentração de BAP (benzilaminopurina) e fotoperíodo. No trabalho publicado por Ridout et al. os dados foram ajustados usando as regressões Poisson, Binomial Negativa, Poisson inflado com zeros (ZIP) e Binomial Negativa inflada com zeros (ZINB).

Para cada distribuição, a variável principal foi explicada pelo fator fotoperíodo (8 ou 16). Na Tabela 6.3.1 é possível comparar as diversas propostas de ajustes para dados com excesso de zeros. A regressão Touchard proporciona um ajuste melhor que as demais propostas, segundo o critério de informação de Akaike (AIC).

Tabela 6.3.1: Ajustes do número de raízes produzidas por 270 brotos no cultivo da maçã.

Tipo	-2Log-veross.	G.L.	AIC
Poisson	1.571,9	268	1.575,9
Binomial Negativa	1.403,9	267	1.409,9
ZIP	1.355,2	267	1.361,2
ZINB	1.336,5	266	1.344,5
Touchard	1.283,2	266	1.289,2

6.4. Touchard x NGPL

Uma das mais recentes generalizações publicadas foi a Nova Poisson-Lindley Generalizada – *New Generalized Poisson-Lindley Distribution* (NGPL), publicada por Bhati et al. [6]. Como apresentado no Capítulo 1.3, a NGPL é obtida assu-

mindo que o parâmetro λ da Poisson segue a distribuição Lindley.

Bhati et al. utilizaram em seu trabalho duas aplicações de contagens em que a distribuição Poisson sabidamente não seria adequada para a modelagem desses dados. Utilizamos esses mesmos exemplos para ilustrar o desempenho da distribuição Touchard.

A Tabela 6.4.1 traz os ajustes dos dados referentes ao número de crises epilépticas pelos modelos Poisson (P), Binomial Negativa (NB), *Generalized Poisson-Lindley* (GPL), *Weighted Generalized Poisson* (WGP) — proposta por Chakraborty [28] — e a Nova Poisson-Lindley Generalizada (NGPL). Foram incluídas nessa comparação as duas versões da distribuição Touchard: a Touchard original (Touch) e a Touchard com três parâmetros (Touch3).

Tabela 6.4.1: Distribuição do número de crises epilépticas

Contagem	Obs.	Esperado						
		P	*NB*	*WGP*	*GPL*	*NGPL*	Touch	Touch3
0	126	74,9	91,0	118,1	121,5	122,0	126,0	125,6
1	80	115,7	86,6	95,8	92,0	91,0	81,0	81,9
2	59	89,3	63,4	59,9	59,0	58,7	58,3	58,2
3	42	46,0	42,6	34,5	35,1	35,2	39,0	38,6
4	24	17,8	27,6	19,2	20,1	20,5	23,4	23,2
5	8	5,5	17,6	10,6	11,2	11,2	12,6	12,6
6	5	1,4	10,5	5,8	6,1	6,4	6,1	6,2
7	4	0,3	6,5	3,2	3,3	3,2	2,7	2,8
8+	3	0,1	5,0	3,9	2,7	2,5	1,7	1,8
Total	351	351	351	351	351	351	351	351
Log-veross.		-636,05	-595,22	-595,83	-594,61	-594,48	-593,05	-593,06
χ^2		256,54	22,53	7,12	5,94	5,75	3,77	3,62
p-valor		$< 0{,}001$	$< 0{,}001$	0,212	0,528	0,562	0,708	0,605

O ajuste da distribuição Touchard aos dados foi bastante satisfatório, obtendo-se uma discrepância pequena entre os dados observados e os valores esperados segundo o modelo, de acordo com o Qui-quadrado. Seu desempenho, para esse conjunto de dados, superou o de todas as distribuições testadas por Bhati et al.

em sua publicação. As duas versões da Touchard proporcionaram uma modelagem bem-sucedida, contudo, o custo da estimação de um terceiro parâmetro faz da versão original da Touchard mais adequada nesse caso.

Outra ilustração utilizada no artigo que propôs a NGPL se refere ao número de acionamentos de seguros de automóvel. O autor comparou, utilizando o teste Qui-quadrado, sua proposta às distribuições Poisson (P), Binomial Negativa (NB), Poisson-Lindley (PL) [19] e Poisson-Lindley Generalizada (GPL). Essa mesma ilustração foi utilizada para avaliar a distribuição Touchard. Na Tabela 6.4.2 pode-se observar que a Touchard se mostrou compatível com as demais propostas de generalização da Poisson, em especial sua versão com três parâmetros.

Tabela 6.4.2: Distribuição do número de acionamentos de seguro de automóvel

Contagem	Obs.	Esperado						
		P	*NB*	*PL*	*GPL*	*NGPL*	Touch2	Touch3
0	1.563	1.544,2	1.566,4	1.569,5	1.566,4	1.564,5	1.564,9	1.564,8
1	271	299,8	261,5	256,3	261,4	264,3	263,0	264,8
2	32	29,1	40,1	41,3	40,2	39,7	40,9	38,3
3	7	1,9	6,0	6,6	6,0	5,6	5,4	5,8
4+	2	0,1	0,9	1,0	0,9	0,9	0,7	1,2
Total	1.875	1.875	1.875	1.875	1.875	1.875	1.875	1.875
χ^2		57,04	3,61	3,87	3,66	3,49	5,07	1,87

6.5. Touchard x Conway-Maxwell-Poisson

A mais recente generalização da Poisson, até a finalização deste trabalho, é a Conway-Maxwell-Poisson Estendida – *Extended Conway-Maxwell-Poisson distribution* (ECOMP), publicada por Chakraborty e Imoto [7]. Como visto no Capítulo 1.3, essa distribuição é uma extensão da clássica distribuição Conway-Maxwell-

Poisson (COMP) publicada em 1962.

Nesta seção, a Touchard será comparada com quatro variações da Conway-Maxwell-Poisson:

- A original COMP;
- A COMNB, uma generalização da Binomial Negativa baseada na COMP [29];
- A GCOMP, generalização da COMP com três parâmetros [30];
- *Extended Conway-Maxwell-Poisson distribution* (ECOMP).

Foram utilizados dois exemplos publicados por Chakraborty e Imoto [7] em seu trabalho sobre a ECOMP: o primeiro se refere a dados sobre o número de infestações do besouro *Dentroctonus frontails* no sudeste do Texas e o segundo exemplo está relacionado ao número de empréstimos por livro na Universidade de Sussex no período de um ano (Falmer, Reino Unido).

Na primeira ilustração não foi obtido um bom ajuste utilizando a distribuição Touchard. Observando os resultados da Tabela 6.5.1, nota-se que a Touchard teve um desempenho melhor do que a clássica COMP, mas inferior a suas generalizações, segundo o teste Qui-quadrado de bondade do ajuste e o AIC.

Já na segunda ilustração, a distribuição Touchard com três parâmetros demonstrou um excelente ajuste aos dados, tendo um desempenho semelhante à ECOMP. A Tabela 6.5.2 traz a comparação entre as distribuições com informações sobre os valores esperados, χ^2 e AIC.

Tabela 6.5.1: Modelos para o número de infestações do besouro *Dentroctonus frontails* no sudeste do Texas.

Contagem	Obs.	Esperado					
		COMP	COMNB	GCOMP	ECOMP	Touch2	Touch3
0	1.169	922,6	1.168,6	1.168,7	1.169,0	1.164,8	1.165,8
1	144	373,4	152,9	151,9	147,6	165,5	163,4
2	92	151,2	80,2	80,7	84,2	72,7	72,7
3	54	61,2	49,8	50,1	52,0	44,9	45,3
4	29	24,8	32,6	32,8	33,2	31,3	31,8
5	18	10,0	21,8	21,8	21,5	22,6	23,0
6	10	4,1	14,7	14,6	14,1	16,3	16,5
7	12	1,6	9,9	9,8	9,4	11,5	11,5
8	6	0,7	6,7	6,6	6,2	7,8	7,8
9	9	0,3	4,5	4,4	4,2	5,1	5,0
10	3	0,1	3,0	2,9	2,8	3,2	3,1
11	2	0,0	2,0	1,9	1,9	1,9	1,9
12	0	0,0	1,3	1,3	1,3	1,1	1,1
13	0	0,0	0,8	0,8	0,8	0,6	0,6
14	1	0,0	0,6	0,6	0,6	0,3	0,3
15	0	0,0	0,4	0,4	0,4	0,2	0,1
16	0	0,0	0,2	0,2	0,3	0,1	0,1
17	0	0,0	0,2	0,2	0,2	0,0	0,0
18	0	0,0	0,1	0,1	0,1	0,0	0,0
19+	1	0,0	0,2	0,2	0,3	0,0	0,0
Total	1.550	1.550	1.550	1.550	1.550	1.550	1.550
χ^2		669,06	8,99	8,65	8,18	16,77	16,51
p-valor		$< 0{,}01$	0,25	0,28	0,23	0,03	0,02
AIC		3.518,85	3.113,24	3.112,99	3.113,84	3.118,72	3.119,67

Tabela 6.5.2: Modelos para o número de empréstimos por livro na Universidade de Sussex no período de um ano (Falmer, Reino Unido)

Contagem	Obs.	Esperado					
		COMP	COMNB	GCOMP	ECOMP	Touch2	Touch3
1	9.647	9.364,8	9.604,4	9.576,2	9.648,6	9.737,9	9.647,4
2	4.351	4.706,6	4.401,3	4.458,5	4.341,4	4.115,3	4.346,4
3	2.275	2.365,5	2.314,7	2.311,9	2.291,3	2.338,1	2.286,2
4	1.250	1.188,9	1.231,2	1.213,9	1.244,9	1.327,7	1.243,8
5	663	597,5	643,2	631,0	663,3	705,9	664,3
6	355	300,3	327,2	322,4	339,5	345,5	340,6
7	154	150,9	161,7	161,6	165,5	155,2	166,0
8	72	75,9	77,7	79,5	76,5	64,2	76,6
9	37	38,1	36,3	38,3	33,5	24,5	33,4
10	14	19,2	16,5	18,2	13,9	8,7	13,8
11	6	9,6	7,3	8,5	5,5	2,9	5,4
12	2	4,8	3,2	3,9	2,1	0,9	2,0
13	0	2,4	1,3	1,8	0,7	0,3	0,7
14+	1	2,5	0,9	1,4	0,6	0,1	0,3
Total	18.827	18.827	18.827	18.827	18.827	18.827	18.827
χ^2		66,83	7,20	14,75	2,37	38,94	2,25
p-valor		$< 0{,}01$	0,51	0,06	0,94	$< 0{,}01$	0,90
AIC		52.472,22	52.414,70	52.422,75	52.411,67	52.440,54	52.406,64

6.6. Dados de futebol: *Premier League*

A *Premier League* é uma liga de futebol profissional disputada por 20 clubes, sendo a principal competição do futebol inglês. O sistema de pontuação é o de pontos corridos, em que cada time disputa 38 jogos, totalizando 380 partidas na temporada. Ao final, os quatro clubes melhor classificados disputam a Liga dos Campeões da UEFA (União das Federações Europeias de Futebol).

Dados sobre as partidas da *Premier League* podem ser obtidos no *site* [1] da organização. Neste exemplo, o interesse está no número de gols ocorridos por partida considerando os jogos "em casa" e os jogos "fora de casa" na temporada 2013-2014.

A distribuição do número de gols em partidas de futebol tem sido explorada considerando as distribuições de Poisson e Binomial Negativa. Aqui, será proposta a distribuição Touchard como alternativa para a modelagem desses dados.

A figura 6.1 traz a distribuição amostral do número de gols marcados. É notável a concentração de zeros e uns nessas distribuições, o que pode levar a uma superdispersão.

A tabela 6.6.1 mostra a distribuição do número de gols observado em comparação com as distribuições Poisson, Binomial Negativa, Touchard e Touchard com três parâmetros (Touchard3).

Note que os valores preditos pelas distribuições Touchard e Touchard3 são muito próximos dos verdadeiros valores observados. Além disso, de acordo com os critérios de informação de Akaike - AIC e Bayesiano - BIC, a Touchard proporciona o melhor ajuste. Ainda com base nesses critérios, a distribuição Touchard se mantém como uma alternativa competitiva aos modelos clássicos Poisson e Binomial

[1] `http://www.premierleague.com`

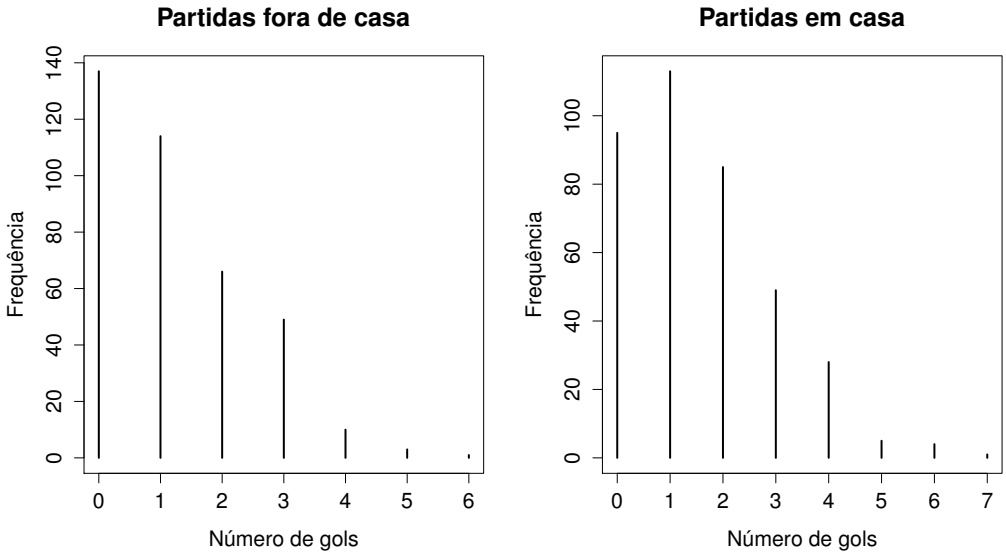

Figura 6.1: Distribuição amostral do número de gols marcados em partidas da *Premier League*.

Negativa.

A tabela 6.6.2 traz as estimativas de médias e variâncias do número de gols dentro e fora de casa, bem como os parâmetros do modelo Touchard obtidos a partir dos dados.

Esses dados confirmam, obviamente, que a média de gols marcados pelos clubes em casa é maior que a média de gols marcados fora de casa. Assim, podemos propor um modelo de regressão considerando como variável explicativa o local onde ocorreu a partida.

Seja y_i o número de gols marcados em uma partida. Assumindo que y_i pode ser modelado pela distribuição Touchard com parâmetros λ_i e δ_i, temos o modelo de regressão em que $\lambda_i = e^{\beta_0+\beta_1 x_i}$ e $\delta_i = \alpha$, no qual x_i assume o valor 1 se o jogo ocorrer fora de casa e 0 se ocorrer na casa do clube. Ou seja, a covariável x explica o parâmetro λ do modelo. Essa escolha é intuitiva, uma vez que a covariável x parece estar mais associada à média ($\bar{y}_{casa} = 1,57$ e $\bar{y}_{fora} = 1,19$) do que à dispersão,

Tabela 6.6.1: Número de gols observado e esperado por partida na *Premier League*.

	# Gols (em casa)								
	0	1	2	3	4	≥ 5	Log-veross.	AIC	BIC
Observado	95	113	85	49	28	10	—	—	—
Touchard	95,0	112,0	86,9	50,1	23,0	13,0	-617,2	1.238,4	1.246,3
Touchard3	95,1	111,7	87,0	50,3	23,1	12,7	-617,2	1.240,4	1.252,3
Bin. Neg.	91,5	118,3	87,0	47,8	21,9	13,5	-617,8	1.239,6	1.247,5
Poisson	78,7	123,9	97,5	51,2	20,1	8,6	-621,3	1.244,6	1.248,5
	# Gols (fora de casa)								
	0	1	2	3	4	≥ 5	Log-veross.	AIC	BIC
Observado	137	114	66	49	10	4	—	—	—
Touchard	134,7	118,0	73,1	34,8	13,4	6,0	-553,9	1.111,8	1.119,7
Touchard3	137,0	111,3	77,0	36,6	13,2	5,0	-553,3	1.112,5	1.124,4
Bin. Neg.	130,5	125,3	72,4	32,7	12,7	6,4	-555,5	1.115,0	1.122,9
Poisson	115,0	137,5	82,1	32,7	9,8	2,9	-559,5	1.121,0	1.124,9

representada pelo coeficiente de variação ($\text{CV}_{casa} = 1,21$ e $\text{CV}_{fora} = 1,21$).

A Tabela 6.6.3 traz as estimativas de máxima verossimilhança para os parâmetros do modelo de regressão Touchard, sendo o desempenho da Touchard comparado à regressão de Poisson e à regressão Binomial Negativa.

Observa-se que a regressão Touchard, de acordo com AIC, apresenta um bom desempenho frente aos modelos clássicos Poisson e Binomial Negativa. Esse resultado ilustra como a Touchard pode ser utilizada como opção para regressão de dados de contagem com superdispersão.

Tabela 6.6.2: Estimativas de máxima verossimilhança dos parâmetros da Touchard, média e variância amostrais do número de gols por partida da *Premier League*.

Local	$\hat{\lambda}$	$\hat{\delta}$	$\hat{\mu}$	$\hat{\sigma}^2$
Fora de casa	2,018725	-1,204640	1,194737	1,444827
Em casa	2,2729176	-0,9454807	1,5736842	1,8969310

Tabela 6.6.3: Estimativas de máxima verossimilhança dos parâmetros da regressão Touchard para o número de gols marcados na *Premier League*.

Regressão	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\alpha}_0$	AIC
Touchard	0,8660	-0,2243	-1,0626	2.349
Bin. Neg.	0,4534	-0,2755	—	2.353
Poisson	0,4534	-0,2755	—	2.366

6.7. Acidentes de trânsito em *NY*

Acidentes de trânsito são modelados tradicionalmente pelas distribuições Poisson e Binomial Negativa. O grande interesse por esse tema pode ser associado à sua alta letalidade: segundo a Organização Mundial de Saúde - OMS, 3.400 pessoas morrem diariamente vítimas de acidentes de trânsito e 10 milhões de pessoas são feridas ou incapacitadas a cada ano. [2]

O portal do estado de Nova Iorque (`https://data.ny.gov`) disponibiliza dados sobre o número de acidentes de trânsito, com informações adicionais sobre o estado da via, condição do tempo, etc. Como ilustração, foram selecionados os dados do distrito de Washington. Como variável de interesse, foi utilizado o número de acidentes ocorridos diariamente nos anos de 2011 a 2013. A figura 6.2 mostra a distribuição do número diário de acidentes de trânsito.

A média diária de acidentes no período foi igual a $2,91$ e a variância igual a $4,06$ acidentes^2, um evidente exemplo da presença de superdispersão nos dados.

A regressão Touchard pode ser tomada como uma nova proposta para modelagem desse tipo de dados. Nesse sentido, a tabela 6.7.1 traz o desempenho da Touchard frente às distribuições Poisson e Binomial Negativa.

Uma análise descritiva mostrou que o número de acidentes está relacionado às estações do ano e aos dias da semana. Foi observado que os dias da semana

[2] `http://www.who.int/violence_injury_prevention/road_traffic/en/`

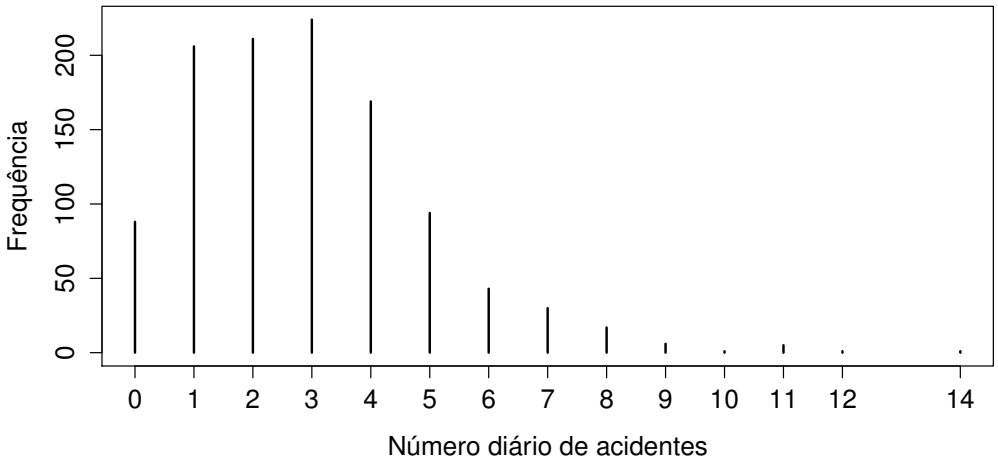

Figura 6.2: Distribuição amostral do número diário de acidentes de trânsito ocorridos em Washington.

sexta-feira e sábado são aqueles em que ocorrem mais acidentes, enquanto que na estação primavera o número de acidentes foi menor.

Desta forma, um modelo proposto para esse exemplo pode ser ajustado com duas covariáveis, sendo $x_1 = 1$ se o dia da semana for sexta-feira ou sábado, e $x_1 = 0$, caso contrário. E ainda, $x_2 = 1$, quando a estação do ano for primavera e $x_2 = 0$, caso contrário. A tabela 6.7.1 traz as estimativas dos parâmetros dos modelos ajustados.

Tabela 6.7.1: Estimativas de máxima verossimilhança dos parâmetros da regressão Touchard para o número de acidentes de trânsito.

Regressão	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\alpha}_0$	AIC
Touchard	1,3609	0,1441	-0,2081	-1,0510	4.425
Bin. Neg.	1,0746	0,1795	-0.2597	—	4.413
Poisson	1,0748	0,1786	-0,2589	—	4.457

Nesse exemplo, segundo o critério de informação de Akaike, a Touchard teve um desempenho um pouco pior que a regressão Binomial Negativa, mas à frente

da regressão de Poisson.

6.8. Dados com subdispersão

Dados com subdispersão são menos comuns que dados com superdispersão. De qualquer maneira, a subdispersão, assim como a superdispersão, pode provocar vieses nas estimativas do modelo ajustado caso se decida pelo uso da distribuição de Poisson.

Um exemplo comum de subdispersão ocorre com dados truncados no zero, ou seja, quando a quantidade zero não é observada na distribuição. Considere, por exemplo, uma variável aleatória Y com distribuição Poisson truncada no zero, sua função de probabilidade por definição é

$$\begin{aligned} P(Y=k) &= P(X=k|X>0) \\ &= \frac{P(X=k)}{1-P(X=0)} \\ &= \frac{e^{-\mu}\mu^y}{y!(1-e^{-\mu})}, \end{aligned}$$

em que $\mu = \mathrm{E}(X)$. O valor esperado e a variância de Y são dados por

$$\mathrm{E}(Y) = \frac{\mu}{1-e^{-\mu}} \quad \text{e} \quad \mathrm{Var}(Y) = \left(\frac{\mu}{1-e^{-\mu}}\right)\left(1-\frac{\mu e^{-\mu}}{1-e^{-\mu}}\right).$$

Note que $\mathrm{Var}(Y) < \mathrm{E}(Y)$.

Como foi visto no Capítulo 2, a distribuição Touchard também é capaz de modelar dados em que a variância é menor do que a média. Para ilustrar esse tipo de ajuste, foram utilizados dados sobre o número de pares de tênis de 60 atletas de

corrida de rua. Esses dados estão disponíveis no livro *Analyzing Categorical Data* [31].

A Tabela 6.8.1 traz o ajuste desses dados pelos modelos Poisson, Poisson truncado no zero e Touchard. Note que a distribuição Poisson produz valores esperados muito distantes dos reais valores observados. Por outro lado, a Poisson truncada no zero e a Touchard proporcionam um ajuste adequado, tendo estas um desempenho muito similar.

Tabela 6.8.1: Número de pares de tênis dos atletas de corrida de rua

Contagem	Obs.	Esperado		
		Poisson	Truncada	Touchard
0	0	5,2	0	0
1	18	12,7	16,8	17,6
2	18	15,5	18,2	18,6
3	12	12,7	13,2	13,0
4	7	7,8	7,2	6,7
5+	5	6,1	4,7	4,1
Total	60	60	60	60
χ^2		8,11	0,22	0,31
p-valor		0,09	0,90	0,86

Capítulo 7

Conclusão

Esta dissertação apresentou a distribuição Touchard: uma generalização da Poisson com dois parâmetros para modelagem de dados não-Poisson. Trouxe ainda uma extensão da Touchard com três parâmetros e uma proposta de modelo linear generalizado assumindo que a variável resposta vem de uma distribuição Touchard.

A flexibilidade em modelar dados com subdispersão, superdispersão e excesso de zeros torna essa distribuição bastante interessante. Além disso, as estimativas dos parâmetros, tanto da distribuição quanto da regressão, são obtidas numericamente a partir de algoritmos simples e com baixo custo computacional.

Para ilustrar a aplicabilidade da distribuição e do modelo de regressão Touchard, dados reais e exemplos de outros artigos foram utilizados para a comparação dos resultados com outras distribuições propostas na literatura para dados de contagem em que a distribuição Poisson não é adequada.

Como observado nas ilustrações, a distribuição Touchard se mostrou competitiva diante dos mais importantes modelos para dados de contagem e, em alguns casos, frente às mais recentes propostas de distribuições para dados de contagem. Mesmo nas situações em que a Touchard foi superada, o ajuste se mostrou bastante

razoável, mantendo-a como uma alternativa para a modelagem dos dados.

A inclusão de um terceiro parâmetro na distribuição Touchard permitiu um melhor ajuste de dados com excesso de zeros. Tal propriedade foi bastante explorada nas aplicações, obtendo-se bons resultados.

Dados resultantes de contagens são frequentes em diversas áreas como saúde, acidentes de trânsito, biologia, controle de qualidade, entres outras. Com base nisso, foi desenvolvido um pacote na linguagem R como forma de atrair pesquisadores das diversas áreas de pesquisa para o uso da Touchard.

Naturalmente, não há a pretensão de colocar a distribuição Touchard como principal opção para a modelagem de dados de contagens. Contudo, ela pode fazer parte da gama de generalizações da Poisson a serem testadas como alternativas na modelagem de dados científicos e de relevância prática.

Em algumas situações, a regressão Touchard, implementada com os algoritmos propostos neste trabalho, falhou. É preciso analisar melhor o comportamento da regressão diante de modelos mais complexos, com fatores com vários níveis e termos de interação. Em trabalhos futuros, a regressão Touchard será melhor explorada.

Também pode ser vista como limitação da Touchard, apesar de suas fórmulas simples, a ausência de formas analíticas fechadas para a média, variância e para os estimadores dos parâmetros da distribuição. Isso dificultou, por exemplo, o uso do *deviance* no diagnóstico de ajuste dos modelos.

Além da revisão da regressão Touchard, deverão ser implementadas novas funções no pacote R, como a Touchard com três parâmetros e seu modelo regressão, bem como o *deviance* como teste de bondade do ajuste utilizando o modelo saturado empírico.

Outros pontos interessantes dessa nova generalização ainda serão explorados: a abordagem bayesiana e o processo Touchard no contexto de processo estocástico

(filtro linear).

Essas pretensões de trabalhos futuros são suportadas pela potencial contribuição que a distribuição Touchard pode trazer à comunidade científica.

Bibliografia

[1] P. C. Consul and G. C. Jain, "A Generalization of the Poisson Distribution," *Technometrics*, vol. 15, no. 4, pp. 791–799, 1973.

[2] B. Efron, "Double exponential families and their use in generalized linear regression," *Journal of the American Statistical Association*, vol. 81, no. 395, pp. 709–721, 1986.

[3] Y. Zou, S. R. Geedipally, and D. Lord, "Evaluating the double Poisson generalized linear model," *Accident Analysis and Prevention*, 2013.

[4] R. W. Conway and W. L. Maxwell, "A queuing model with state dependent service rates," *Journal of Industrial Engineering*, vol. 12, pp. 132–136, 1962.

[5] G. Shmueli, T. P. Minka, J. B. Kadane, S. Borle, and P. Boatwright, "A useful distribution for fitting discrete data: revival of the Conway-Maxwell-Poisson distribution," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 54, pp. 127–142, jan 2005.

[6] D. Bhati, D. Sastry, and P. M. Qadri, "A New Generalized Poisson-Lindley Distribution: Applications and Properties," *Austrian Journal of Statistics*, vol. 44, no. 4, p. 35, 2015.

[7] S. Chakraborty and T. Imoto, "Extended Conway-Maxwell-Poisson distribution and its properties and applications," *Journal of Statistical Distributions and Applications*, no. 2014, pp. 1–14, 2016.

[8] D. Lambert, "Zero-Inflated Poisson Regression, With an Application To Defects in Manufacturing," *Technometrics*, vol. 34, no. 1, pp. 1–14, 1992.

[9] R. Matsushita, D. Pianto, B. B. D. Andrade, and A. Cançado, "The Touchard Distribution and Process." Não publicado, 2015.

[10] F. A. Haight, *Handbook of the Poisson distribution*. Wiley, 1967.

[11] S. M. Stigler, *The history of statistics: The measurement of uncertainty before 1900*. Harvard University Press, 1986.

[12] J. A. Nelder and R. W. M. Wedderburn, "Generalized linear models," *Journal of the Royal Statistical Society. Series A (General)*, vol. 135, no. 3, pp. 370–384, 1972.

[13] Student, "An Explanation of Deviations from Poisson's Law in Practice," *Biometrika*, vol. 12, no. 3-4, pp. 211–215, 1919.

[14] M. Greenwood and G. U. Yule, "An Inquiry into the Nature of Frequency Distributions Representative of Multiple Happenings with Particular Reference to the Occurrence of Multiple Attacks of Disease or of Repeated Accidents," *Journal of the Royal Statistical Society Series B*, vol. 83, no. 2, pp. 255–279, 1920.

[15] Y. Qu, G. J. Beck, and G. W. Williams, "Polya-Eggenberger Distribution: Parameter Estimation and Hypothesis Tests," *Biometrical Journal*, vol. 32, no. 2, pp. 229–242, 1990.

[16] C. Bliss and R. Fisher, "Fitting the Negative Binomial Distribution to Biological Data / Note on the efficient fitting of the Negative Binomial," 1953.

[17] F. E. Satterthwaite, "Generalized Poisson Distribution," *The Annals of Mathematical Statistics*, vol. 13, no. 4, pp. 410–417, 1942.

[18] E. C. Maceda, "On the Compound and Generalized Poisson Distributions," *The Annals of Mathematical Statistics*, vol. 19, no. 3, pp. 414–416, 1948.

[19] M. Sankaran, "The Discrete Poisson-Lindley Distribution," *Biometrics*, vol. 26, no. 1, pp. 145–149, 1970.

[20] R. Shanker, S. Sharma, and R. Shanker, "A Two-Parameter Lindley Distribution for Modeling Waiting and Survival Times Data," *Applied Mathematics*, vol. 4, no. February, pp. 363–368, 2013.

[21] J. Touchard, "Sur les cycles des substitutions," *Acta Mathematica*, vol. 70, no. 1, pp. 243–297, 1939.

[22] C. A. Charalambides and O. D. Chrysaphinou, "Partition polynomials in fluctuation theory," *Mathematische Nachrichten*, vol. 106, no. 1, pp. 89–100, 1982.

[23] O. V. Kuz'min and O. V. Leonova, "Touchard polynomials and their applications," *Diskretnaya Matematika*, vol. 12, no. 3, pp. 60–71, 2000.

[24] R. Matsushita, "The three-parameter Touchard distribution and its applications." Não publicado, 2016.

[25] J. DeLeeuw, "Introduction to Akaike (1973) Information Theory and an Extension of the Maximum Likelihood Principle," 1992.

[26] B. F. Thorndike, "Applications of Poisson ' s Probability Summation," vol. 2, no. 1, pp. 95–113, 1923.

[27] M. Ridout, C. G. Demétrio, and J. Hinde, "Models for count data with many zeros," *International Biometric Conference*, no. December, pp. 1–13, 1998.

[28] S. Chakraborty, "On some distributional properties of the family of weighted generalized Poisson distribution," *Communications in Statistics—Theory and Methods*, vol. 39, no. 15, pp. 2767–2788, 2010.

[29] S. Chakraborty and S. H. Ong, "A COM-Poisson-type generalization of the negative binomial distribution," *Communications in Statistics - Theory and Methods*, vol. 45, no. 14, pp. 4117–4135, 2016.

[30] T. Imoto, "A generalized Conway–Maxwell–Poisson distribution which includes the negative binomial distribution," *Applied Mathematics and Computation*, vol. 247, pp. 824–834, 2014.

[31] J. S. Simonoff, *Analyzing Categorical Data*. Springer Texts in Statistics, Springer New York, 2013.